Грибач Т.А. (Минск, МГЛУ)

ВОПРОСЫ ГЕНЕРАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Генерация текстов на естественном языке находится по-прежнему в области приоритетных задач современной компьютерной лингвистики.

До сегодняшнего дня основными технологиями порождения текстов компьютером являлись шаблонные технологии и лингвистически мотивированные технологии [1]. Шаблонные технологии достаточно просты, надежны и находят широкое промышленное применение. Самые простые шаблонные системы используют готовые фрагменты написанного человеком текста без их дополнительной обработки. Более сложные системы позволяют задавать отдельные грамматические компоненты текста или комбинировать шаблонные высказывания и, таким образом, получать связный текст, используя при этом определенные лексические и грамматические знания о языке. Порожденные тексты выглядят вполне естественно, т.к. представляют собой последовательность фрагментов готового текста. Однако необходимо отметить, что подобные системы работают только с очень жесткими типами текстов. Использование шаблонов при порождении текста компьютером эффективно и целесообразно тогда, когда заранее известна структура порождаемого текста и существует возможность заранее определить его лексическое наполнение. Системы, созданные на основе лингвистически мотивированных технологий являются более значимыми, потому что позволяют генерировать тексты с относительно свободным

содержанием, которое не может быть задано при помощи готовых фрагментов текста. В данном случае источником содержания порождаемого текста являются данные, представленные в виде баз данных, баз знаний или выражений на формализованных языках. Тип входных данных не всегда предсказывает тип выходного текста, поэтому последний определяется извне пользователем. Подобные системы создают тексты одного типа, но в разных предметных областях, или в одной предметной области, но на разных языках. Современные системы порождения текста состоят из следующих основных компонентов: оболочки, планировщика и лингвистического реализатора. Оболочка определяет назначение системы генерации и характер базы знаний, на основе которой происходит построение текста. Она выполняет две основные функции: инициирует процесс порождения и обуславливает цели, которые должны быть достигнуты в результате синтеза высказывания. Планировщик определяет пути достижения поставленных целей в данном предметном контексте. Он обеспечивает: 1) выбор информации, которая должна быть представлена в тексте; 2) определение того, как должна быть представлена эта информация; 3) выбор способа взаимодействия с лингвистическими данными. В частности, планировщик проводит структурирование текста, построение синтаксической структуры предложений и выбор соответствующей лексики. Опираясь на концептуальное представление текста, выработанное планировщиком, лингвистический реализатор осуществляет окончательный контроль за процессом порождения текста. Он принимает все окончательные морфологические и синтаксические решения и отвечает за грамматически правильное оформление текста. В процессе генерации текста компьютером ученые выделяют три относительно независимых этапа: 1) макропланирование (построение плана текста); 2) микропланирование (построение плана каждого предложения текста); 3) языковое оформление предложений средствами конкретного языка [1]; [2]; [3].

Одним из последних решений в области автоматической обработки текста является система понимания, анализа и перевода текстов на естественных языках ABBYY Compreno.

Технология представляет собой лингвистическую платформу, позволяющую решать различные прикладные задачи относительно компьютерного анализа текста на более высоком уровне. Сюда относятся задачи в области перевода, интеллектуального поиска на одном или нескольких языках, выявления в текстах ключевых объектов, фактов и связей между ними [4]; [5].

Особенность технологии ABBYY Compreno заключается в том, что текст представляется компьютеру не просто последовательностью слов, а в виде «осмысленной» информации, что даёт новые возможности для её дальнейшей обработки.

В основе этой технологии лежит универсальная иерархия понятий и модель отношений между этими понятиями (т.е. иерархия универсальных семантических значений и отношений между ними). Существует множе-

ство различных языков, но так или иначе они описывают схожую систему понятий. Эта система образует семантическое дерево понятий, в англоязычном варианте обозначается как USH (Universal Sematic Hierarchy) [6].

Не менее важной частью технологии является полный синтаксический разбор текста. Синтаксис — это способ «кодирования» семантических отношений в конкретном языке. Сами семантические отношения универсальны, а способы их реализации в каждом языке — свои. В каких-то языках установлен линейный порядок, в других используются падежи, предлоги, специальные служебные слова, где-то используется всё сразу. Для каждого языка синтаксическое описание делается заново, но сами средства, которые разные языки используют для кодирования смысла, перечислимы. При описывании нового языка используются разные элементы конструктора (тот же линейный порядок, различные типы синтаксических преобразований, грамматические значения, предлоги, специальные конструкции).

Технология Compreno также успешно определяет и более сложные синтаксические связи, такие как: анафора, эллипсис. Выделяемые системой связи между понятиями также выражаются в древесной структуре, фактически передают смысл написанного, и несут важную информацию для поиска или перевода. Таким образом, система стремится к определению смысла текста, написанного на обычном языке, позволяя машине «понять» этот текст и трансформировать его в универсальное представление, не зависящее от языка.

Используя семантическое дерево понятий, синтаксическое описание языка, а также статистику взаимоотношений между словами, технология Compreno производит полный анализ текста и при переводе его на другой язык использует слова, соответствующие правильным ветвям дерева понятий и отношениям, выявленным при разборе исходного предложения.

ЛИТЕРАТУРА

1. Болдасов, М.В. Парадигмы генерации ЕЯ текстов в инструментальной среде DEMLING. Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2003». — М., 2003. — С. 66–75.

2. Соколова, Е.Г. Генерация текстов на естественном языке — состояние вопроса и прикладные системы. Сер. 2. Информационные процессы и системы № 10, 2005. — С. 12–22.

3. Бусел, Т.В. О лингвистически мотивированных подходах к проблеме генерации текстов на естественном языке / Вестн. МГЛУ. Сер.1, Филология. — 2009. № 2. — С. 170– 178.

4. Онтоинженер: от сотворения мира к порождению сущностей. [Электронный ресурс]. — Режим доступа: http://habrahabr.ru/company/abbyy/blog/246039/ — Дата доступа: 10.03.2015.

5. ABBYY Intelligent Search SDK. [Электронный ресурс] — Режим доступа: http://www.abbyy.ru/isearch/compreno/ — Дата доступа: 09.03.2015.

6. Научные разработки в бизнесе. [Электронный ресурс] — Режим доступа: http://www.abbyy.ru/science/technologies/business/compreno/ — Дата доступа: 10.03.2015.